Federal Aviation
Administration

DOT/FAA/AM-12/19
Office of Aerospace Medicine
Washington, DC 20591

Development, Validation, and Fairness of a Biographical Data Questionnaire for the Air Traffic Control Specialist Occupation

Michelle Dean
San Diego State University
San Diego, CA 92127

Dana Broach
Civil Aerospace Medical Institute
Federal Aviation Administration
Oklahoma City, OK 73125

December 2012

Final Report

NOTICE

Technical Report Documentation Page

1. Report No.
DOT/FAA/AM-12/19

2. Government Accession No.

3. Recipient's Catalog No.

4. Title and Subtitle
Development, Validation, and Fairness of a Biographical Data Questionnaire for the Air Traffic Control Specialist (ATCS) Occupation

5. Report Date
December 2012

6. Performing Organization Code

7. Author(s)
Dean MA,[1] Broach DM[2]

8. Performing Organization Report No.

9. Performing Organization Name and Address
[1]San Diego State University, San Diego, CA 92127
[2]FAA Civil Aerospace Medical Institute, P.O. Box 25082
Oklahoma City, OK 73125

10. Work Unit No. (TRAIS)

11. Contract or Grant No.

12. Sponsoring Agency name and Address
Office of Aerospace Medicine
Federal Aviation Administration
800 Independence Ave., S.W.
Washington, DC 20591

13. Type of Report and Period Covered

14. Sponsoring Agency Code

15. Supplemental Notes
This work was completed under approved FAA Human Factors research task 1209AC082110.HRR523.AV9300, Evaluation of ATCS Biographical Data and Interview Selection Procedures

16. Abstract

Development and validation of a biographical data (“biodata”) instrument for selection into the Air Traffic Control Specialist occupation is described. Bootstrapping was used to estimate correlations between item responses to the Applicant Background Assessment (ABA; 142 items; n=266), Biographical Questionnaire (BQ; 145 items; n=482), and average supervisory job performance ratings. Scoring keys were developed for the most predictive 80, 100, and 120 items from the instruments. Reliabilities for the proposed scales ranged from .74 to .78. Criterion-related validities were .59, .62, and .63 for the 80-, 100-, and 120-item versions, respectively.

Each version of the biodata scale had significant incremental validity over the AT-SAT composite score, accounting for 29% to 32% additional variance in average job performance ratings. Score distributions and cut-scores by race and sex were investigated. Differences (*d*) in mean scores by gender and ethnicity were generally low (ranging from -.08 to .37). Finally, cut score analyses were performed to examine pass rates of demographic subgroups using banding and percentiles.

Based on the findings of this study, it is recommended that the 80-item biodata scale, renamed the Controller Background Assessment Survey (CBAS), be further developed as a potential ATCS selection procedure.

17. Key Words
Air Traffic Control Specialist, Personnel Selection, Biographical Data, Bootstrap

18. Distribution Statement
Document is available to the public through the Internet:
www.faa.gov/go/oamtechreports

19. Security Classif. (of this report)
Unclassified

20. Security Classif. (of this page)
Unclassified

21. No. of Pages
16

22. Price

ACKNOWLEDGMENTS

Research reported in this paper was conducted under the Air Traffic Program Directive/Level of Effort Agreement between the Human Factors Research and Engineering Group (AJP-61), FAA Headquarters, and the Aerospace Human Factors Division (AAM-500) of the FAA Civil Aerospace Medical Institute.

The opinions expressed are those of the authors alone, and do not necessarily reflect those of the Federal Aviation Administration, the Department of Transportation, or federal government of the United States of America.

CONTENTS

Development, Validation, and Fairness of a Biographical Data Questionnaire for the Air Traffic Control Specialist (ATCS) Occupation

The measurement of biographical data (or "biodata") encompasses the notion of asking individuals to recall and report their typical, and sometimes, specific behaviors or experiences in a referent situation, generally from an earlier time in their lives (Mumford & Owens, 1987; Nickels, 1994). While different approaches such as diaries have been used to collect biodata, the most common form is that of a scale, survey, inventory, or questionnaire. Such biodata instruments have demonstrated reasonable and useful reliability and validity in the prediction of job performance in a variety of occupations (see Stokes, Mumford, & Owens, 1994). Average cross-validities in the .3 to .4 range have been reported for biodata selection instruments in narrative and meta-analytic review (Asher, 1972; Hunter & Hunter, 1984; Reilly & Chao, 1982; Schmitt, Gooding, Noe, & Kirsh, 1984). Moreover, biodata scales can be constructed so as to have less adverse impact by race without significant loss in criterion-related validities (Dean, 1999).

The United States federal government has long had an interest in the development and validation of biodata, reaching back to World War I, at least (see Farmer, 2002, for a review). The Federal Aviation Administration (FAA) has conducted similar investigations of the air traffic control specialist (ATCS) occupation. Following the 1981 controller strike, the FAA faced an enormous organizational challenge in rebuilding this highly technical workforce (see Broach, 1998, 2005). While the core of the post-strike ATCS selection process from 1981 through the mid-1990s was a cognitive aptitude test battery, researchers at the FAA's Civil Aerospace Medical Institute (CAMI) investigated alternative assessments, including biodata. Two instruments in particular were administered to several thousand newly hired air traffic controllers for research purposes between 1981 and 1992: the Applicant Background Assessment (ABA) and the Biographical Questionnaire (BQ). Research with these instruments indicated that biodata had promise as a personnel selection tool for the ATCS occupation.

The technical approach to scoring these instruments was straightforward and traditional: responses to specific items such as prior experience, self-reported grades in high school math, and the candidate's expectations about future performance as an ATCS, were regressed on criteria representing performance in training (Broach, 1992; Cobb, Young, & Rizutti, 1976; Collins, Manning, & Taylor, 1984; Collins, Nye, & Manning, 1992; Taylor, VanDeventer, Collins, & Boone, 1983). Alternatively, item responses were regressed on training performance criteria to empirically derive a scale (Broach, 2008). Both approaches fall under the broad rubric of empirical keying, that is, developing the scoring key on the basis of item-criterion relationships. However, empirical keys can capitalize on chance associations and, hence, could yield unstable estimates of item and scale validities.

Cross-validation is the primary strategy for countering capitalization on chance or sample-specific associations in the development of an empirically-keyed scale (e.g., a scale based on data rather than judgment). Cross-validation ensures estimates of item and scale validities are 1) not simply capitalizing on chance characteristics of the calibration sample and 2) appropriately attenuated to reflect the effect of random sampling error in any future samples (e.g., future applicant pools). A second strategy is to use very large samples for development. As the sample size increases, scoring weights become more stable and prediction errors are smaller. As sample size decreases, the weights are less stable and errors are larger. A third strategy is to hold out a portion of the original sample for cross-validation. However, this approach reduces confidence in the scoring weights (Cohen, Cohen, West, & Aiken, 2003; Murphy, 1983). There is no generally acceptable minimum standard sample size required for cross-validation, but Gatewood, Field, and Barrick (2008) recommended a sample size of at least 300 for cross-validating biodata instruments. Assuming a 3:1 ratio for the development to cross-validation sample sizes, about 1,200 cases might be required to develop and cross-validate an empirically-keyed biodata scale.

The FAA enjoyed the luxury of large samples after the 1981 ATCS strike as the basis for development of empirically-keyed biodata scales. However, the veracity of biodata item responses, particularly in high-stakes selection processes, has long been a concern with empirically keyed scales (Lautenschlager, 1994). The stakes are certainly high for the FAA's ATCS selection process. ATCS job applicants are generally highly motivated. For example, some invest thousands of dollars in tuition and fees to attend two- and four-year colleges participating in the FAA's Air Traffic Control Collegiate Training Initiative (ATC-CTI) with the hope of getting hired by the FAA. They seek out and share information about the selection procedures in a number of on-line forums. The pay-off is the prestige and pay associated with the job. The stakes are equally high for the agency. ATCS training is intensive, extensive, and expensive. Completion of all training phases takes an average of two to three years, depending on facility assignment (FAA, 2011; Manning, 1998). Failures waste FAA training dollars and personnel resources. They also result in fewer people becoming controllers, a critical concern for the agency as the post-strike generation of controllers reaches retirement age (FAA, 2011). Given the high stakes, it is reasonable to expect that applicants will attempt to answer questions about life experiences, attitudes, and expectations in what they believe is an employer-desired direction. At the same time, it is reasonable – and necessary – for the agency to counter that tendency to gain honest and accurate information about an applicant's job-related life experiences, attitudes, and expectations as the basis for justifiable and accurate employment decisions.

An alternative to binary scoring ("item keying") is to score each response option. In "response option keying," each item response option is analyzed separately as if it were an item in and of itself. A response option contributes to the overall score if and only if it is correlated significantly with the criterion (Kluger, Reilly, & Russell, 1991). While still an empirical scoring strategy, response option keying is thought to be less susceptible to response biases (Kluger et al., 1991). However, response-option keying can be problematic because of the sheer number of correlations to be computed. Capitalization on chance characteristics of the sample is possible with small samples relative to the number of computed item response-criterion correlations.

One possible solution to this problem is bootstrapping. Bootstrapping is a statistical technique that holds promise for statistically analyzing small sample sizes. Bootstrapping estimates the sampling distribution of a statistic (e.g., correlation between to variables X and Y, or r_{xy}) by iteratively resampling cases from a set of observed data. Basically, *B* "bootstrap" samples of size *N* are taken with replacement from the original sample of size *N*, then the statistic of interest (in this case, a correlation between an individual response option and the criterion of interest) is generated within each bootstrap sample and then saved to a file. An investigation using *B*=1,000 bootstrap samples of size *N* is able to approximate the actual sampling distribution that would have been obtained if multiple independent samples of size *N* were drawn from the population (Efron & Tibshirani, 1993). Bootstrapping is computationally time-intensive as the sample at hand is resampled with replacement many times to derive a statistic of interest. Bootstrap estimation uses all available information in the sample in estimating prediction error; therefore, a hold-out cross-validation sample is not needed. An additional benefit of bootstrap estimation is that data need not meet the usual parametric assumptions (i.e., the data need not be normally distributed; Efron & Tibshirani). Using bootstrap analyses on CAMI archival controller data, Russell, Dean, and Broach (2000) found that sample sizes as small as 175-200 incumbents might provide sufficient data for accurately estimating the "true" population validity coefficient.

These three concepts – empirical keying, response-option scoring, and bootstrapping – were utilized in this study to derive a biodata instrument that might be used in the selection of air traffic controllers. The study was conducted in three steps. First, bootstrap analyses were conducted with datasets for the BQ and ABA independently. The ABA and BQ items were rank-ordered by their average correlation with the criterion, and the top 80, 100, and 120 items selected for inclusion in a biodata scale. Scale scores were then computed for each of the proposed scales. Second, hierarchical regression analyses were conducted to determine the incremental validity of the 80-, 100-, and 120-item biodata scales over the computerized Air Traffic Selection and Training (AT-SAT) aptitude test battery composite score in predicting the supervisory ratings criterion. Third, analyses of score distributions and score banding were conducted to assess fairness of the proposed scale.

METHOD

Sample

Development and validation. ABA, BQ, and criterion data were extracted from AT-SAT concurrent, criterion-related validation database. Overall, 1,232 incumbent controllers participated in the AT-SAT validation (Keenan, 2001, p. 31). The AT-SAT validation database included 266 records with ABA and criterion data and 482 with BQ and criterion data. The bootstrap analysis was run for each instrument separately to develop an empirical response-option scoring key for that instrument. Previous work on bootstrapping demonstrated that these sample sizes were sufficient to estimate the true population validity coefficient for each response option (Russell et al., 2000). The participants in the AT-SAT validation study were not selected on the basis of their ABA or BQ responses, so incidental restriction in range[1] (on those instruments) was likely. However, there is no accepted procedure for correcting the bootstrapped item-criterion correlations for incidental restriction in range. Finally, the bootstrap procedure handled each item individually, without regard to or dependence on other item responses. There are no data suggesting that item responses in one instrument were dependent on responses to the other instrument.

Table 1
Development and validation sample demographic information (from the AT-SAT concurrent, criterion-related validation)

	AT-SAT ATCS (*N*=1,232)		ABA & Criterion (*n*=266)		BQ & Criterion (*n*=482)		ABA, BQ, & Criterion (*n*=260)	
	N	%	N	%	N	%	N	%
Males	909	72.2	226	85.0	404	83.4	221	85.0
Females	178	19.5	40	15.0	78	16.2	39	15.0
Missing data	145							
Native American	77	6.3	1	0.4	3	0.6	1	0.4
Asian/Pacific Islander	9	0.7	4	1.5	6	1.2	4	1.5
Black	95	7.7	23	8.6	42	8.8	22	8.5
Hispanic	64	5.2	11	4.1	26	5.4	11	4.2
Non-minority	804	65.3	227	85.0	404	83.8	221	85.0
Other	20	24.6	1	0.4	1	0.2	1	0.4
Mixed race	4	0.3						
Missing data	159	12.9						

A subset of 260 controllers in the AT-SAT dataset had full and complete ABA, BQ, and criterion data (i.e., 260 of the 482 cases with BQ also had ABA data). Data from this subset (*n*=260) were used to estimate the criterion-related and incremental validity of the response-option scored 80-, 100-, and 120-item scales. As shown in Table 1, the development and validation samples were largely male (~85%) and White (~85%).

Group differences. The sample used to investigate group differences was drawn from the CAMI archival database on the 27,925 controllers hired between 1981 and 1992 ("post-strike" controllers). There were 5,826 records in CAMI's archival database with both ABA and BQ data for the period 1988 through 1992. These 5,826 records included the 260 cases in the validation sample (complete BQ, ABA, & criterion data). However, as the AT-SAT concurrent, criterion-related validation study records made available for this research were de-identified, a definitive match was not possible between the CAMI archival and the AT-SAT validation data. As in the development and validation samples, the CAMI archival sample was largely male (81.3%) and White (90.6%). Blacks (*n*=253; 4.5%) and Hispanics (*n*=179; 3.2%) were numerically the largest minority groups in the archival sample.

Measures

Predictors. The Applicant Background Assessment (ABA) is a 142-item multiple choice (five response options per item) biodata questionnaire. The ABA was based on: 1) a review of ATCS occupational qualification standards, 2) a review of job analyses conducted by the FAA, 3) a review of previous biodata research done at the FAA, 4) interviews with training staff members to determine characteristics that differentiated those controllers who succeeded or failed in training, and 5) interviews with ATCS supervisors to ascertain characteristics differentiating good and poor ATCSs. The items included on the ABA were limited to those dealing with experiences that were under the control of the applicant.

The Biographical Questionnaire (BQ) is a 145-item inventory that was developed based on items from Owens' Biographical Questionnaire (Owens & Schoenfeldt, 1979). The BQ items tap eight areas: 1) educational background, 2) prior military or civilian experience in ATC, 3) importance placed on various factors (e.g., salary, benefits, job security), 4) time expected to become an effective ATCS, 5) commitment to an ATCS career, 6) work-related attitudes, 7) expected satisfaction with aspects of ATCS careers, and 8) general personal information (e.g., socioeconomic status growing up, alcohol and tobacco usage; Collins, et al, 1992).

AT-SAT is a computerized selection test battery designed for ATCS selection. This test battery was designed to replace the two-stage selection process in which ATCS applicants completed an Office of Personnel Management (OPM) test battery and then a nine-week training program at the FAA Academy. This multiple-hurdle selection process, used from late 1981 through early 1992, was both time consuming and expensive (Broach, 1998; Ramos, Heil, & Manning, 2001a). AT-SAT was developed based on the results of the Separation and Control Hiring Assessment (SACHA) job analysis, an extensive analysis of the ATCS job (Nickels, Bobko, Blair, Sands, & Tartak, 1995). Specifically, the worker requirements determined necessary for the job of ATCS were used to design a series of computerized tests to assess these worker requirements. The overall AT-SAT composite score, as computed in the 1997-1998 concurrent, criterion-related validation study, was used as the baseline predictor in this study.

Job performance criterion. Two criterion measures were developed in the course of the AT-SAT concurrent validation study: a computer-based measure of technical performance; and a job performance rating. The computer-based performance measure (CBPM) was designed as a practical and economical assessment of a controller's technical proficiency in separating aircraft (Hanson, Borman, Mogilka, Manning, & Hedge, 1999). The ratings-based measure consisted of over-the-shoulder ratings used by peers and supervisors to assess typical on-the-job performance (Borman, Hedge, Hanson, Bruskiewicz, Mogilka, Manning, et al., 2001). The assessment consists of behaviorally anchored rating scales (BARS) for the ten performance categories identified as important to the ATCS occupation by subject matter experts: 1) maintaining safe and efficient air traffic flow; 2) maintaining attention and vigilance; 3) prioritizing communicating, and informing; 4) coordinating; 5) managing multiple tasks; 6) reacting to stress; 7) adaptability and flexibility; 8) technical knowledge; 9) teamwork; and 10) overall effectiveness. A composite of the two criteria was used in the concurrent validation of AT-SAT (Ramos, Heil, & Manning, 2001b). In subsequent analyses, the average rating of the ten BARS was used as the criterion (Wise, Tsacoumis, Waugh, Putka, & Hom, 2001). Therefore, to be consistent with the later validity and reweighting studies of AT-SAT, the average rating of job performance was used as the criterion for estimating the validity of ABA and BQ item responses and the overall 80-, 100-, and 120-item scales.

Procedures

The first step was to convert the biodata item-level data into response option-level data. Converting the ABA from item-level to response option-level data resulted in 710 response options for the ABA (142 items with five response options each) and 725 response options for the BQ (145 items with five response options each). The majority of the BQ items contained five response options with the exception of 26 items that had either three or four response options. For programming simplicity, all BQ items were set to have five response options. For example, for those items with four response options, the "5th option" created by the response option program would simply be scored "0" across all cases and would not contribute to the biodata scores. Data preparations also involved screening for no variance response options (i.e., no one or very few [1 or 2] controllers chose that response option) as these response options would cause the bootstrap program to stop and need to be restarted (and it is already known that these response options would receive a "0" weight in the scoring key due to lack of variance). Response options with no variance were manually assigned a "0" weight.

Next, the predictor data were standardized in the validation dataset (the ratings criterion data were already standardized). Standardization was performed to simplify and more efficiently run the bootstrap programs. SYSTAT 10.2 (Systat Inc., Chicago, IL), the statistical package used to perform the bootstrap procedure, can more efficiently save regression coefficients using its bootstrap routine in the Regression Procedure, compared to saving correlations generated in its Correlation Procedure. In a bivariate analysis, the standardized regression coefficient (β) is equal to the familiar correlation coefficient (r_{xy}).

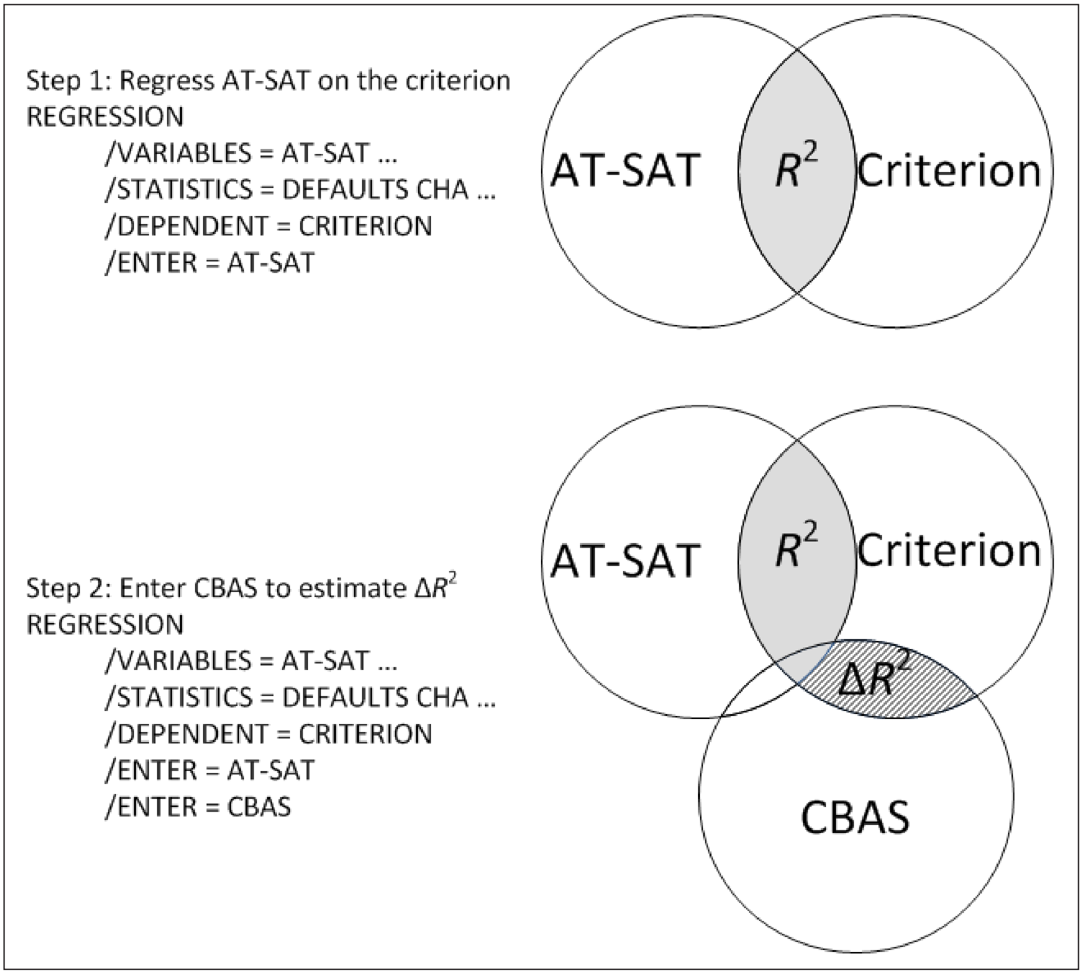
The next step in preparing the data for bootstrapping was to identify subsets of the data with complete cases on each of the biodata instruments and the criterion. One subset included only cases that had complete data on both the ABA and the criterion (*n*=266), and another subset included cases that had complete data on both the BQ and the criterion (*n*=482). This helped to decrease the time required to run each bootstrap.

The last step was to execute the bootstrap analyses for each subset of data to estimate the correlation between response options and the criterion. Each bootstrap program generated 1,000 samples of the same size as the original sample from which it was drawn, with replacement. Correlations between each respective biodata response option and the criterion (one correlation for each of the 1,000 random samples taken) were generated and saved to a file.

The output from the bootstrap analysis (1,000 correlations between each biodata response option and average supervisory rating of ATCS job performance) was used to develop the scoring weights for the biodata instruments. The average of 1,000 correlations was used to weight each response option in the scoring keys. The larger the correlation, the stronger the relationship of that particular response option with the criterion. Three scoring keys were developed using the 80, 100, and 120 most predictive items from the ABA and BQ. The sum of the absolute values of the response options' correlations with the criterion (within item) was used to identify the ABA and BQ items with the most predictive set of response options. Each scoring key started with an arbitrary point of 100, then adding or subtracting the weights of response options chosen, then multiplying the overall score by 10 to increase the range of scores.

Analyses

The biodata scores generated from the 80-, 100-, and 120-item scoring keys were correlated with the job performance measure to obtain criterion-related validities in the validation dataset (e.g., the records with full and



Step 1: Regress AT-SAT on the criterion

```
REGRESSION
        /VARIABLES = AT-SAT ...
        /STATISTICS = DEFAULTS CHA ...
        /DEPENDENT = CRITERION
        /ENTER = AT-SAT
```

Step 2: Enter CBAS to estimate ΔR^2

```
REGRESSION
        /VARIABLES = AT-SAT ...
        /STATISTICS = DEFAULTS CHA ...
        /DEPENDENT = CRITERION
        /ENTER = AT-SAT
        /ENTER = CBAS
```

Figure 1
Incremental validity analysis (with SPSS syntax)

complete data on the ABA, BQ, and criterion; n=260). Second, the incremental validity of the biodata score was estimated through hierarchical regression. In hierarchical regression analysis, the variance in the dependent variable is uniquely partitioned based on the order in which the (correlated) independent variables are entered (Cohen et al., 2003, p. 158). The standardized regression weight (β) for the independent variable entered in the first step is equal to its zero-order correlation with the criterion. In the second step, the effect of the second independent variable is estimated, taking into account the first predictor (see Cohen et al. p. 67). The critical question in incremental validity is the additional variance (change in the multiple correlation coefficient, or ΔR^2) in the criterion explained by the second predictor (Hunsley & Meyer, 2003; see Figure 1).

Score distributions by ethnicity and sex, cut scores, and subgroup pass rates using score banding and percentile scores were investigated. The number of minorities in the subset of AT-SAT cases with ABA, BQ, and criterion data (n=260) was very small, making fairness analyses with the validation sample impractical. However, the larger archival sample of 5,826 controllers hired between 1988 and 1992 was diverse enough to support the analysis of biodata scale scores by subgroup. Two analyses were conducted, both requiring comparisons between groups. First, subgroup differences (*d*s) were calculated for the 80-, 100-, and 120-item biodata scales in the larger archival data set (n=5,826) for cases with complete data on demographics and biodata. A *d* statistic represents the difference between group means (e.g., Male vs. Female) divided by the pooled (sample-weighted, within-group) standard deviation of the two groups (Hunter & Schmidt, 2004). Second, score bands were created by using ±2 times the standard error of measurement for each of the biodata scales (Cascio, Outtz, Goldstein & Zedeck, 1991). Cut scores were also examined by percentile.

RESULTS

Statistical Analyses

Descriptive statistics for the top 80-, 100-, and 120-item biodata scales, AT-SAT predictor composite, and the job performance criterion are found in Table 2. The mean scores for the 80-, 100- and 120-item scales were very nearly the same while the score variance increased slightly with the number of items. Some items were reverse-coded for the purposes of calculating reliability estimates so that all items were scored in the same direction. Reliabilities (Cronbach's α) for the 80-, 100-, and 120-item scales were .74, .74, and .78 respectively. Correlations among the study variables are also reported in Table 2. The correlations between AT-SAT and the proposed biodata scale scores were .37, .34, and .34. These correlations suggested that AT-SAT and biodata might be tapping different controller personal characteristics. The correlation between scores on the AT-SAT test battery and the average job performance rating in this analysis was .26 (n=260, p<.01), compared to .21 reported by Ramos, et al. (2001b, Table 5.5.1).

The zero-order correlations between the 80-, 100-, and 120-item biodata scales and average job performance ratings (e.g., "criterion-related validities") were .59, .62, and .63 respectively. The results of the hierarchical regression analysis are presented in Table 3 for the 80-, 100-, and 120-item biodata scales. As shown in Table 3, AT-SAT was a valid predictor of average job performance ratings (β=.26, p < .01; R^2=.07, p < .01) in the first step of the incremental validity analyses. The biodata scales demonstrated incremental validity over AT-SAT. The standardized regression coefficient for the 80-item biodata scale was .58; it accounted for an additional 29% of variance in the criterion (ΔR^2=.29, p < .01). The 100- and 120-item versions accounting for an additional 32% of criterion variance as shown in Table 3.

Fairness Analysis

The first step in the fairness analysis was to investigate group differences in mean scores on the proposed scales. The scoring keys for the 80-, 100-, and 120-item scales were applied to the CAMI archival data (n=5,826). Mean group differences (d) were calculated for three

Table 2
Descriptive statistics and intercorrelations for predictors and criterion in validation dataset[1]

	Measure	Mean	SD	1	2	3	4	5
1	80-item Biodata Scale	107.12	15.78	.74[2]				
2	100-item Biodata Scale	107.33	17.24	.99	.74			
3	120-item Biodata Scale	108.56	19.14	.98	.98	.78		
4	AT-SAT	73.39	7.96	.37	.34	.34	-	
5	Job Performance	5.02	.73	.59	.62	.63	.26	-

Notes: [1] All correlations are significant at p < .01; n=260

[2] Reliabilities for biodata scales on diagonal (all items). (Reliabilities calculated without categorical items for the 80, 100, and 120-item scales were .78 [65 items], .79 [80 items], and .83 [97 items], respectively).

Table 3
Hierarchical regression analyses

		80-items			100-items			120-items		
Step	Predictor	β	ΔR^2	R^2	β	ΔR^2	R^2	β	ΔR^2	R^2
1	AT-SAT Score	.26**	.07**	.07**	.26**	.07**	.07**	.26**	.07**	.07**
2	Biodata	.58**	.29**	.35**	.53**	.32**	.39**	.61**	.32**	.39**

**p<.01

comparisons: female-to-male (d_{F-M}); Black-to-White (d_{B-W}); and Hispanic-to-White (d_{H-W}). There were too few American Indian/Alaskan Native and Asian/Pacific Islander trainees to support meaningful comparisons. Results from the mean group differences analysis are presented in Table 4. The d_{F-M} was quite small across all of the biodata scales in the archival dataset and much smaller than reported for AT-SAT (d_{F-M}=.44; Ramos et al., 2001b). The mean group differences on the proposed biodata scales were were larger by race, with d_{B-W} =.37. and d_{H-W} =.22 for the 80-item scale. The d_{B-W} for the biodata scales was about half of the d_{B-W} for Blacks on AT-SAT. However, the d_{H-W} for the biodata scales was comparable to the d_{H-W} reported by Waugh (2001) for AT-SAT.

The second step in the fairness analysis was to examine pass rates by subgroup. Two sets of subgroup pass rate analyses were conducted. First, score banding was used to develop groupings of scores. Scores within a band are treated as equivalent (Henle, 2004). In this instance, bandwidth was calculated as two times the standard error of measurement. Three ranges of scores were created (with the lowest band being made up of the lowest two bands combined) to model category ranking as used in the FAA. Next, selection ratios by group were computed in two score banding scenarios: selection from the highest band only (the functional equivalent of selecting from the "Well Qualified" band only); and selection from both the highest and middle bands (selecting from the "Well Qualified" and "Qualified" bands). Overall, less than 10% of any group was placed in the highest score band (Figure 2 for 80-item scale; results were comparable for the 100- and 120-item scales[2]).

Table 4
Biodata mean gender and ethnic group differences (*d*)[1]

	80 Items	100 Items	120 Items	AT-SAT[2]
Male-Female *d*	-.08	-.07	-.15	-.44
Black-White *d*	.37	.39	.41	.74
Hispanic-White *d*	.22	.22	.21	.24

Notes: [1]Based on CAMI archival data: d_{F-M} n=5,826 (n_{Female}=1,087); d_{B-W} n=5,382 (n_{Black}=253); d_{H-W} n=5,308 ($n_{Hispanic}$=179)
[2]From Table 5.6.5, Ramos, Heil & Manning, 2001b

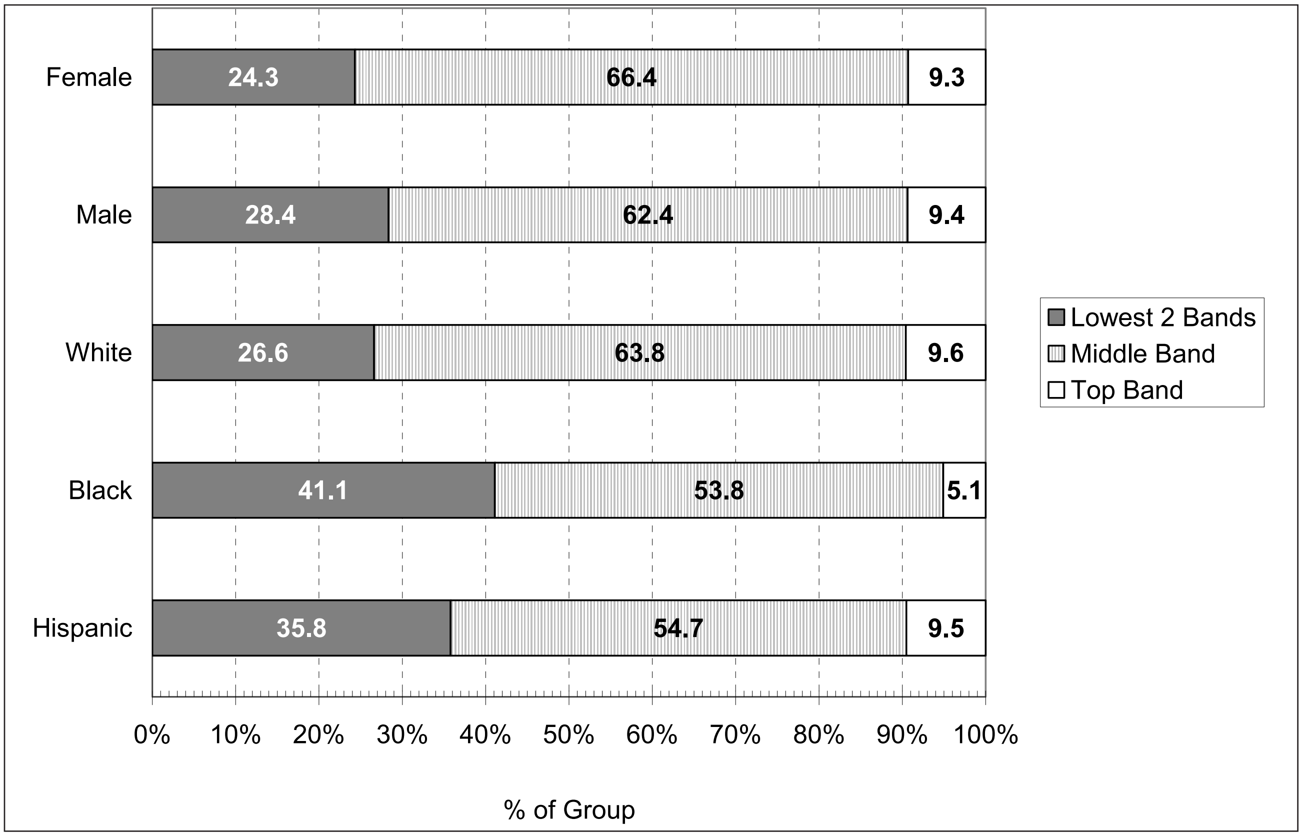


Figure 2
Score band replacements by group for 80-item biodata instrument

The first selection scenario was stringent, with selections from the top score band only. In this stringent scenario, the female-to-male selection ratio was .99 for the 80-item version. The Black-to-White selection ratio was .53 for the 80 item version. When selections were made from the top score band only, the ratios of Hispanic-to-White selection rates were .99 for 80-item version. These results suggest that, for the 80-item version, selection from the top band would not adversely impact women or Hispanic applicants.

The second selection scenario was relatively lenient, with selection from the top and middle score bands. About three-quarters of men, women, and Whites were in the middle and top score bands. However, as shown in Table 5, only about 65% of Black and Hispanics were in the middle and top score bands. Selection from the middle and top bands resulted in a female-to-male selection ratio at or above 1.00 for all versions of the biodata instrument. The Black-to-White selection ratio was .80 for the 80-item version of the biodata instrument but less than .80 for the longer versions. Selection from the middle and top bands combined resulted Hispanic-to-White selection ratio greater than .80. The results for the 80-item version indicate that selection from the middle and top score bands combined would not adversely impact women, Blacks, or Hispanics.

The final step in the fairness analysis was to investigate pass rates by percentile cut scores. Selection ratios were computed at each percentile. The female-to-male ($SR_{\text{F-M}}$), Black-to-White ($SR_{\text{B-W}}$), and Hispanic-to-White ($SR_{\text{H-W}}$) selection ratios are plotted by percentile in Figure 3 for the 80-item version. The $SR_{\text{H-W}}$ and $SR_{\text{H-W}}$ selection ratios were greater than .80 at all percentiles. However, the $SR_{\text{B-W}}$ was less than .80 at higher percentile scores. For example, selection at the 80th percentile (20% selected, 80% rejected) resulted in $SR_{\text{B-W}}$ of .61. Setting a cut score

Table 5
Cut score results for 80-, 100-, and 120-item biodata scale by number and percentage of each group within-band for CAMI archival data[1]

	Male		Female		Black		Hispanic		White	
	N	%	*N*	%	*N*	%	*N*	%	*N*	%
					80 Items					
Top Band	447	9.4	101	9.3	13	5.1	17	9.5	495	9.6
Middle Band	2,955	62.4	722	66.4	136	53.8	98	54.7	3,271	63.8
Lowest 2 Bands	1,337	28.4	264	24.3	104	41.1	64	35.8	1,363	26.6

[1]CAMI archival data (*N*=5,826): M=4,739, F=1,087; B=253, H=179, W=5,129

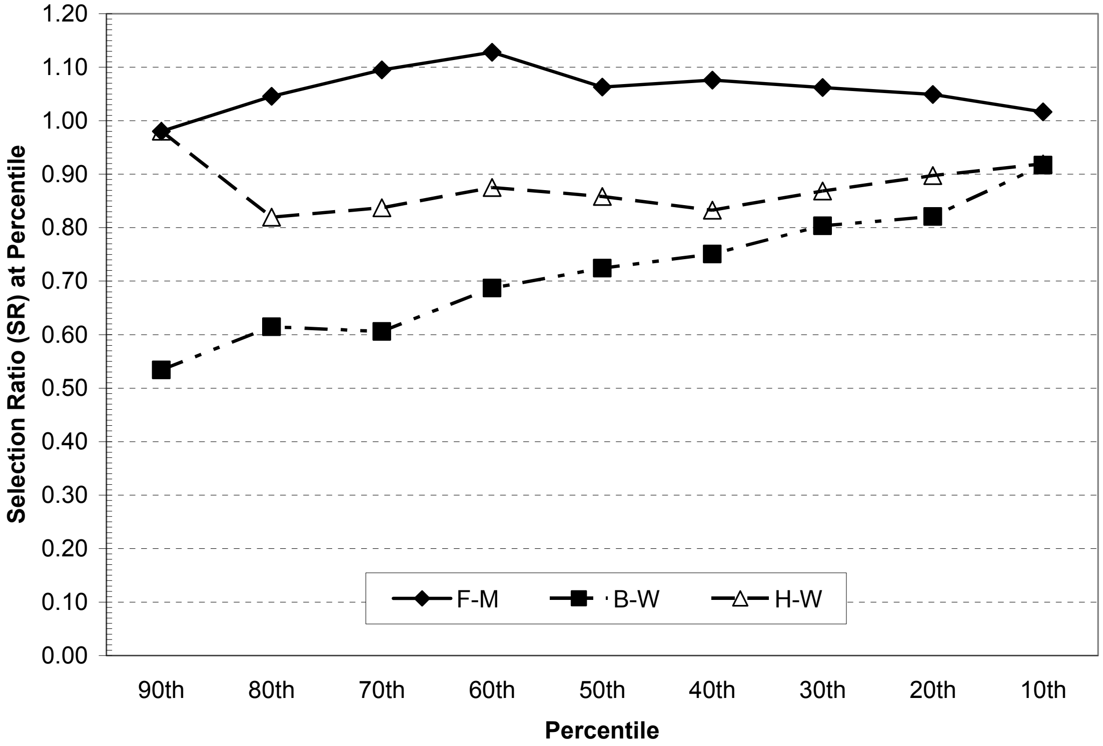


Figure 3
Selection ratios by group for the 80-item biodata instrument

at the 50th percentile (50% selected, 50% rejected), for example, resulted in SR_{B-W} of .72 for the 80-item version, still below the .80 rule-of-thumb promulgated in the *Uniform Guidelines for Employee Selection Procedures*. In other words, selection on higher percentile scores on the 80-item version was unlikely to adversely impact women and Hispanics but adverse impact on Black candidates was possible.

DISCUSSION

The results of this study suggest that an empirically-keyed, response option-scored biodata instrument has validity as a predictor of ATCS job performance ratings. From a test fairness perspective, biodata yielded nearly identical mean scores across gender and ethnicity scores that were well below *d*s typically found for tests of general mental ability—which tend to yield high subgroup differences of around 1.0. The results of this study follow the typical research findings on biodata—that it holds promise for prediction while at the same time causing less adverse impact potential relative to tests of general cognitive ability. Finally, the 80-item version was more efficient (fewer questions for about the same statistical gain) than either the 100- or 120-item versions of the biodata scale.

Three additional studies are recommended. First, an investigation of the internal structure of the biodata instrument is recommended. This investigation should include the convergent and discriminant validity of the biodata scale with personality measures. Second, an assessment of the incremental validity of biodata in predicting more objective measures of controller job performance is warranted. Third, cross-validation of the biodata scale on the incoming generation of air traffic controllers is recommended. Example job performance criteria include performance in initial occupational training at the FAA Academy, typically in the first few months after hire. More distal criteria include on-the-job training performance assessments and achieving Certified Professional Controller (CPC) at the first assigned field facility (1-3 years from hire (FAA, 2011, p. 40)). Further investigations of CBAS in relation to these outcomes seem warranted as data become available and empirical studies are technically feasible under the relevant professional guidelines, standards, principles, and practices for the validation of employee selection procedures.

REFERENCES

Asher, J.J. (1972). The biographical item: Can it be improved? *Personnel Psychology*, *25*, 251–269.

Borman, W.C., Hedge, J.W., Hanson, M.A., Bruskiewicz, K.T., Mogilka, H., Manning, C., … & Horgen, K.E. (2001). Development of criterion measures of air traffic controller performance. In R.A. Ramos, M.A. Heil, & C.A. Manning (Eds.) *Documentation of validity for the AT-SAT computerized test battery: Volume II* (pp. 1–12). (FAA Report No. DOT/FAA/AM-01/6). Washington, DC: Federal Aviation Administration, Office of Aviation Medicine.

Broach, D. (1992, June). *Non-cognitive predictors of performance in radar-based air traffic control training.* Paper presented at the 4th Annual Convention of the American Psychological Society, San Diego, CA.

Broach, D. (Ed.) (1998). *Recovery of the FAA air traffic control specialist workforce, 1981-1992.* (DOT/FAA/AM-98/23). Washington, DC: Federal Aviation Administration Office of Aviation Medicine.

Broach, D. (2005). A singular success: Air traffic control specialist selection 1981-1992. In B. Kirwan, M. Rodgers & D. Schafer (Eds.). *Human factors impacts in air traffic management* (p. 177–205). Burlington, VT: Ashgate.

Broach, D. (2008). *Overview of CAMI Life Experiences Questionnaire (Version 1.0).* Unpublished manuscript, FAA Civil Aerospace Medical Institute Aerospace Human Factors Research Division (AAM-500).

Cascio, W.F., Outtz, J., Zedeck, S., & Goldstein, I.L. (1991). Statistical implications of six methods of test score use in personnel selection. *Human Performance*, *4*, 233–264.

Cobb, B.B., Young, C.L., & Rizutti, B.L. (1976). *Education as a factor in the selection of air traffic controller trainees.* (FAA Report No. DOT/FAA/AM-76/16). Washington, DC: Federal Aviation Administration Office of Aviation Medicine.

Cohen, J., Cohen, P., West, S.G. & Aiken, L.S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Hillsdale, NJ: Erlbaum.

Collins, W.E., Manning, C.A., & Taylor, D.K. (1984). A comparison of prestrike and poststrike ATCS: Biographic factors associated with Academy training success. In A.D. VanDeventer, W.E. Collins, C.A. Manning, D.K. Taylor, & N.E. Baxter (Eds.). *Studies of poststrike air traffic control specialist trainees: I. Age, biographical factors, and selection test performance related to Academy training success* (pp. 7–14). (FAA Report No. DOT/FAA/AM- 84/6). Washington, DC: Federal Aviation Administration Office of Aviation Medicine.

Collins, W.E., Nye, L.G., & Manning, C.A. (1992). Poststrike air traffic control trainees: Biodemographic predictors of success in selection and screening. *International Journal of Aviation Psychology*, *2* 213-223.

Dean, M.A. (1999, June). *A response-option examination of biodata adverse impact and criterion validity.* Presented at the 1999 International Personnel Management Association Assessment Council meeting, St. Petersburg, FL.

Efron, B., & Tibshirani R.J. (1993). *An Introduction to the bootstrap*. New York: Chapman & Hall.

Farmer, W.L. (2002). *Characteristics of biodata keys as a function of scaling method, sample size, and criterion.* Dissertation Abstracts International: Section B: The Sciences and Engineering, 63(3-B).

Federal Aviation Administration. (2011). *A plan for the future: 10-Year strategy for the air traffic control workforce 2011–2020.* Washington, DC: Federal Aviation Administration. Last retrieved May 25, 2011 from http://www.faa.gov/air_traffic/publications/controller_staffing/media/CWP_2011.pdf

Gatewood, R.D., Feild, H.S., & Barrick, M. (2008). *Human resource selection*. (6th ed.)Fort Worth, TX: Harcourt.

Hanson, M.A., Borman, W.C., Mogilka, H.J., Manning, C., & Hedge, J.W. (1999). Computerized assessment of skill for a highly technical job. In F. Drasgow & J.B. Olson-Buchanan (Eds.), *Innovations in computerized assessment* (pp.197–220). Mahwah, NJ: Erlbaum.

Henle, C.A. (2004). Case review of the legal status of banding. *Human Performance*, *17*, 415–432.

Hunsley, J. & Meyer, G.J. (2003). The incremental validity of psychological testing and assessment: Conceptual, methodological, and statistical issues. *Psychological Assessment*, *15*, 446–455.

Hunter, J.E., & Hunter, R.F. (1984). Validity and utility of alternative predictors of job performance. *Psychological Bulletin, 96*, 72–98.

Hunter, J.E., & Schmidt, F.L. (2004). *Methods of meta-analysis: Correcting for error and bias in research findings* (2nd Ed.). Newbury Park, CA: Sage.

Keenan, P.A. (2001). *Biographical and computer experience information: Demographics for the validation study.* In R.A. Ramos, M.A. Heil, & C.A. Manning (Eds.) Documentation of validity for the AT-SAT computerized test battery: Volume II (pp. 31–35). (FAA Report No. DOT/FAA/AM-01/6). Washington, DC: Federal Aviation Administration, Office of Aviation Medicine.

Kluger, A.N., Reilly, R.R., & Russell, C.J. (1991). Faking biodata tests: Are option-keyed instruments more resistant? *Journal of Applied Psychology*, 76, 889–896.

Lautenschlager, G.J. (1994). Accuracy and faking of background data. In G.S. Stokes, M.D. Mumford, & W.A. Owens, (Eds.), *Biodata handbook: Theory, research and use of biographical information in selection and performance prediction* (pp. 391–419). Palo Alto, CA: Consulting Psychologists Press.

Manning, C.A. (1998). Air traffic control specialist field training programs, 1981-1992. In D. Broach (Ed.), *Recovery of the FAA Air Traffic Control Specialist Workforce, 1981-1992* (pp. 23–32). (FAA Report No. DOT/FAA/AM-98/23). Washington, DC: Federal Aviation Administration Office of Aviation Medicine.

Mumford, M.D., & Owens W.A. (1987). Methodological review: Principles, procedures, and findings in the application of background data measures. *Applied Psychological Measurement*, *11*, 1–31.

Murphy, K.R. (1983). Fooling yourself with cross-validation: Single sample designs. *Personnel Psychology, 36,* 111–118.

Nickels, B.J. (1994). The nature of biodata. In G.S. Stokes, M.D. Mumford, & W.A. Owens, (Eds.), *Biodata handbook: Theory, research and use of biographical information in selection and performance prediction* (pp. 1–16). Palo Alto, CA: Consulting Psychologists Press.

Nickels, B.J., Bobko, P., Blair, M.D., Sands, W.A., & Tartak, E.L. (1995). *Separation and control hiring assessment (SACHA) final job analysis report* (Deliverable item 007A under FAA contract DFTA01-91-C-00032). Washington, DC: Federal Aviation Administration, Office of Personnel Management.

Owens, W.A., & Schoenfeldt, L.F. (1979). Toward a classification of persons. *Journal of Applied Psychology*, *53*, 569-607.

Ramos, R.A., Heil, M.A., & Manning, C.A. (2001a). *Documentation of validity for the AT-SAT computerized test battery: Volume I*. (FAA Report No. DOT/FAA/AM-01/5). Washington, DC: Federal Aviation Administration, Office of Aviation Medicine.

Ramos, R.A., Heil, M.A., & Manning, C.A. (2001b). *Documentation of validity for the AT-SAT computerized test battery, Volume II*. (FAA Report No. DOT/FAA/AM-01/6). Washington, DC: Federal Aviation Administration Office of Aerospace Medicine.

Reilly, R.R., & Chao, G.T. (1982). Validity and fairness of some alternative employee selection procedures. *Personnel Psychology*, 35, 1–62.

Russell, C.J., Dean, M.A., & Broach, D. (2000). *Guidelines for bootstrapping validity coefficients in ATCS selection*. (FAA Report No. DOT/FAA/AM-00/15). Washington, DC: Federal Aviation Administration, Office of Aviation Medicine.

Schmitt, N., Gooding, R.Z., Noe, R.A., & Kirsch, M. (1984). Meta-analyses of validity studies published between 1964 and 1982 and the investigation of study characteristics. *Personnel Psychology, 37,* 407-422.

Stokes, G.S., Mumford, M.D., & Owens, W.A. (1994). *Biodata handbook: Theory, research and use of biographical information in selection and performance prediction*. Palo Alto, CA: Consulting Psychologists Press.

Taylor, D.K., VanDeventer, A.D., Collins, W.E. & Boone, J.O. (1983). Some biographical factors associated with success of air traffic control specialist trainees at the FAA Academy during 1980. In A.D. VanDeventer, D.K. Taylor, W.E. Collins, & J.O. Boone (Eds.). *Three studies of biographical factors associated with success in air traffic control specialist screening/training at the FAA Academy* (pp. 6–11). (FAA Report No. DOT/FAA/AM-83/6). Washington, DC: Federal Aviation Administration Office of Aviation Medicine.

Waugh, G. (2001). Predictor-criterion analyses. In R.A. Ramos, M.A. Heil, & C.A. Manning (Eds.) *Documentation of validity for the AT-SAT computerized test battery: Volume II (pp. 37–42).* (FAA Report No. DOT/FAA/AM-01/6). Washington, DC: Federal Aviation Administration, Office of Aviation Medicine.

Wise, L.L., Tsacoumis, S., Waugh, G.W., Putka, D.J., & Hom, I. (2001). *Revision of the AT-SAT.* (Unpublished Manuscript). Washington, DC: Federal Aviation Administration Office of the Assistant Administrator for Human Resources Management.

Endnotes

[1] Incidental restriction in range occurs when a sample is selected on the basis of a different variable than the one being analyzed. For example, the controllers in this analysis were selected on the basis of completing the FAA Academy and field on-the-job training; as a consequence, their scores on the proposed biodata scales might have been incidentally restricted. See personnel selection texts such as Gatewood, Feild, & Barrick, 2008.

[2] Analyses for the 100- and 120-item scales can be requested from the second author.